II. GENE CHARACTERIZATION PANEL

Design of Gene Characterization Studies: an Overview

Duncan C. Thomas

This collection of papers from the Gene Characterization Panel addresses design issues in studies aimed at assessing the population characteristics of cloned genes, such as their allele frequencies, penetrance, variation in these parameters across subpopulations, and gene–environment and gene–gene interactions. This paper provides an overview of the various designs that have been suggested, including cohort and case–control designs using independent and related individuals as well as optimal multistage sampling and hybrid designs. Various statistical (bias and efficiency) and practical considerations are suggested for evaluation of the alternative designs, with the aim of posing the question, “What is the optimal design for a particular situation”? The answer to this question clearly depends on such contextual issues as nature of the outcome variable, the gene frequency and genetic relative risk, and the importance of gene–environment and gene–gene interactions. Further methodologic work might be usefully directed toward assessment of the seriousness of the population stratification problem in general as well as methods of dealing with it, the utility of registries of high-risk families, and the merits of various hybrid designs for gene discovery and gene characterization. [Monogr Natl Cancer Inst 1999;26:17–23]

Aims of Characterization Studies

The class of “gene characterization studies” includes those studies involving a measured gene that is either known to be causally related to some trait or for which such an association is to be tested. Thus, studies of linkage or association with marker genes, or both, that are not thought to be causal factors but merely tools for mapping causal genes are the subject of the Gene Discovery Panel in this volume.

Before proceeding further, it is worth pausing to clarify the meaning of the word “causal” in the genetic context. A gene is a causal risk factor for a trait if the expected outcome depends on an individual’s genotype at that locus, all other risk factors (genetic and environmental) being held constant. For a dichotomous disease trait, the expected outcome is the age-specific (and all-other-factors-specific) incidence rate, otherwise known as “penetrance.” Thus, a gene is causal if this penetrance varies by genotype, after adjusting for all other causal factors. In principle, this definition would exclude associations with marker genes that are due entirely to population associations with other causal genes, but in practice the definition is hard to operationalize, since it might require adjusting for causal loci that have not yet been identified or that are in complete linkage disequilibrium with the truly causal locus. But at a minimum, a locus must be functionally significant to be considered causal.

In the early stages of an investigation, one might wish to test for an association with a candidate gene that has been suggested as a possible causal agent, perhaps by virtue of being located in a region suggested by linkage studies, perhaps by virtue of knowledge of some metabolic function that is plausibly related to the disease process, or perhaps by homology with a known causal gene in other species. In statistical terms, we are interested in testing the null hypothesis that the genetic relative risk associated with this gene is unity. Of course, rejection of the statistical null hypothesis does not by itself establish causality; such a judgment still depends on other criteria familiar to epidemiologists, such as absence of confounding (especially by other genes) or bias in the study design, biologic plausibility (e.g., a functional pathway), consistency across studies, and concordance with other types of knowledge.

Once a gene has been cloned and its causal connection with the disease has been reasonably well established, many questions remain about its quantitative aspects. These aspects can be broadly grouped into two categories: population characteristics and risk characteristics. Under the heading of population characteristics, we are primarily interested in the prevalence of alleles and the variation in prevalence with various population characteristics, such as ethnicity or other genes with which it is in disequilibrium. Under the heading of risk characteristics are all aspects of the dependence of penetrance on genotype. Foremost among these is the main effect of the gene that might be summarized in terms of a genetic relative risk (the ratio of age-specific incidence rates) or the absolute penetrance functions themselves, including the mode of inheritance (dominant, additive, recessive, or codominant). Next, one might be interested in the joint effects of the gene with other genes (G × G interactions), with environmental risk factors (G × E interactions), or the modifying effect of age or other host factors; such interactive effects are arguably most naturally tested in terms of relative risks, but ultimately it is the entire age-specific penetrance function with all its modifiers that we seek to estimate. Finally, we might be interested in the population attributable risk and in the attributable risk in various subgroups (e.g., by family history or ethnic group), quantities that involve both gene frequency and penetrance.

Each of these objectives must be studied in a particular context, depending on what is already known about the gene in question. This context will determine the feasibility and appropriateness of various design options. Some aspects of context include the following:

1) The nature of the outcome (quantitative, dichotomous, or age at onset); in studies of cancer, the end point is most appropriately viewed as censored age-at-onset data.
2) Mode of inheritance.
3) Whether the gene has a single mutant allele or is highly polymorphic.
4) Whether the mutant allele(s) are rare or common.

Correspondence to: Duncan C. Thomas, Ph.D., Department of Preventive Medicine, University of Southern California, 1540 Alcazar St., CHP-220, Los Angeles, CA 90033-9987 (e-mail: dthomas@rcf.usc.edu).

© Oxford University Press
5) Whether the main effect of the gene is strong or weak.
6) Whether there are other known genes or environmental factors with strong effects.
7) Whether there is a single genetic hypothesis under study or a large number of loci or even a genome-wide association scan.

**Available Study Designs**

We begin with the basic premise that estimation of population parameters (any of those discussed above, such as penetrance or allele frequency or their modifiers) requires some form of population-based study. Of course, depending on the sampling design, this may lead to study samples that are not representative of the entire population (e.g., samples of families with multiple cases), but the samples must ultimately be referable to a population base by means of a well-characterized sampling process if unbiased estimates of population parameters and valid tests of statistical hypotheses are to be possible. With this proviso, a wide range of study designs are available for addressing these aims, but most can be viewed as variants of the usual epidemiologic cohort and case–control designs for studying disease risks in relation to measurable risk factors. Zhao et al. (1,2) provide an overview of alternative study designs and their corresponding likelihoods.

For studying rare diseases (like cancer), there are compelling advantages to some form of case–control design, since a general population cohort study would require a very large cohort or long periods of follow-up, or both, to accrue sufficient cases. However, many of the genes of interest (like BRCA1) are also rare, so that an unselected series of cases may not yield an adequate number of mutation carriers. These two considerations suggest that neither the standard case–control nor the cohort designs will be optimal, and some form of multistage sampling will be called for. Furthermore, some of the more interesting designs will involve families, requiring consideration of the subtleties of dependent data, unlike the usual epidemiologic studies of independent individuals.

**Case–Control Designs**

The basic case–control design involves a comparison between case patients with the disease under study and control subjects who are free of the disease (at least at the time the case occurred) in terms of possibly causal factors (here the gene under study and its possible modifiers [host factors, other genes, and environmental factors]). Case patients and control subjects are frequently matched on confounding factors (such as age, sex, and race) that are not of primary interest. The two basic variants of this design are distinguished by their choice of control subjects: randomly selected from the source population of the case patient (the general population or subpopulations thereof) or family members of the case patient. Among family members, siblings or cousins are natural choices because control subjects should be roughly matched on age if environmental factors are also under consideration. An alternative involves the use of parents, not as control subjects per se but as a means of generating the distribution of possible genotypes the case patient could have inherited from the parents (what we shall call below “pseudosibs”). The relative merits of these alternatives are discussed by Witte et al. (3) and in the papers by Caporaso et al. (4) and Gauderman et al. (5) in this volume.

**Cohort Designs**

Whereas the case–control methodology is inherently limited to study of a particular disease, the cohort methodology allows direct observation of all of the end points that may be associated with a particular genetic syndrome, a particularly important consideration for genes whose effects may be pleiotropic. The need for long-term follow-up for chronic diseases like cancer can sometimes be alleviated by using a retrospective cohort study design, but for genetic studies, this study design would require availability of stored biologic samples for genotyping. As discussed by Langholz et al. (6) in this volume, more than one million individuals with stored samples have been enrolled in several large cohort studies dating back one or two generations that provide a rich resource for gene characterization studies. Most cohort studies of genetic factors to date have, however, used a familial cohort design, as discussed by Gail et al. (7) in this volume.

In a case–control study of familial aggregation, one might obtain family history information on case patients and control subjects and treat this information as a risk factor on which to compare case patients and unrelated population control subjects. Conversely, it is also possible to view the family members as a cohort at risk, classified in terms of their “exposure” to the disease status of the sampled case patient or control subject (whom we shall hereafter call the “proband,” irrespective of whether affected or not). In a gene characterization study, the comparison of family members’ outcomes is made in terms of measured genotypes, either their own or the proband’s. The primary distinctions between the design alternatives are whether the family members’ genotypes are also available and how the probands have been sampled.

Perhaps the simplest design is a study of disease incidence in first-degree relatives of genotyped probands as a cohort study. Wacholder et al. (8) use the term “kin-cohort design,” and Gail et al. (9) use the term “genotyped proband design.” Here, the proband might be either affected or unaffected, but if the mutation is rare, affected probands will have a higher yield of mutation carriers. [If the set of probands comprises both cases and unrelated population controls, this design is sometimes called a “case–control family study” (1,10), although it is analyzed as a cohort study.] In this design, the relatives themselves are not genotyped, but the incidence is compared between relatives of carriers and noncarriers. Because roughly 50% of first-degree relatives of carriers will also be carriers and relatives of noncarriers will have roughly the population carrier frequency, it is possible to estimate the age-specific penetrance functions for a dominant gene from such a comparison (9,11,12). Analogous calculations for a recessive gene could also be devised.

This basic design could be extended to include more distant relatives, using maximum likelihood methods to fit a form of segregation analysis model involving penetrance and allele frequency parameters, conditional on the proband’s genotype and mode of ascertainment (9). Allowance for the residual dependencies between family members because of sharing of other genetic or environmental risk factors is necessary to avoid bias (7) and obtain valid confidence intervals. A more direct estimate of penetrance is possible if family members’ genotypes can also be obtained (7,9). The cohort members can then be compared in terms of their own genotypes rather than their probands’ genotypes and penetrance estimated using elementary methods of cohort analysis (except for the allowance for residual familiality
to obtain valid confidence intervals). More research is needed about the optimal selection of additional family members to genotype in these designs—whether it is better to genotype affected or unaffected family members, parents or siblings, near or distant relatives, for example; these issues are discussed in this volume in the context of the National Cancer Institute’s Cooperative Family Registry for Colorectal Cancer Research (13,14).

If a familial cohort study is to involve collection of extended families, then it is essential that one either select a fixed pedigree structure in advance (e.g., ascertain all family members of probands out to second degree) or follow the rules of sequential sampling of pedigrees (15) to avoid bias. Specifically, at any stage of the pedigree-building process, it is legitimate to use the data that have already been collected to decide whether to continue extending the pedigree, and it is also legitimate to use knowledge of the structure of the family at risk in a proposed branch of the family. However, use of anecdotal information about the existence of additional cases in that branch would lead to upwardly biased estimates of penetrance and allele frequencies. Furthermore, all of the data that are collected in this systematic fashion must then be included in the analysis (irrespective of whether additional cases were found). These rules were established for the purpose of segregation analysis, but they are equally applicable to familial cohort studies aimed at characterizing measured genes.

**Multistage Sampling**

Whether using a case–control or a cohort approach, the statistical efficiency of a study may be substantially improved by sampling families in two or more stages, based on a surrogate for genotype like family history (13,16). For example, Whittemore and Halpern (16) describe a three-stage sampling design for a case–control family study of prostate cancer. In the first stage, case patients and population control subjects are selected and asked about the prevalence of prostate cancer in their fathers and brothers; in the second stage, a sample of case patients and control subjects with and without a family history (FH) are subsampled and extended-family histories are obtained and medically confirmed; families with three or more cases are then entered into the third-stage sample from which blood samples are drawn for genotyping. The question then arises as to how to select the sampling fractions at the second stage to minimize the variance of the parameter estimates (penetrances or allele frequency) or to maximize the yield of carrier families for a linkage analysis. For this design and a particular choice of parameters, they show that the optimal choice for parameter estimation is to sample 100% of case FH+ families, 34%–64% of control FH+ families, 12%–20% of case FH– families, and 3%–13% of control FH– families. A similar approach can be taken to optimizing the design of case–control studies with population control subjects (17) or family control subjects (5) or familial cohort studies with only the proband genotyped. For example, the efficiency of familial case–control designs for estimating the main effect of a genotype or a $G \times E$ interaction effect can be substantially improved by restriction to multiple-case families in various ways.

A limitation of multistage sampling is that it tends to focus on only a single disease outcome, whereas many cancer-predisposing conditions cause multiple cancers to occur in families. A sampling design that is optimized to select families that would be most informative for a given cancer may then be inefficient for studying these other end points associated with the same genes. This issue also arises in the analysis as a need to account for competing risks, as in the analysis of breast and ovarian cancers by Whittemore et al. (12). Finally, any comparisons of relative efficiency must carefully account for the different costs at each of the stages of sampling.

**Use of High-Risk Families**

Pedigrees with many cases are highly informative for linkage analysis but typically have not been ascertained in any population-based manner. Frequently, such families have initially come to the attention of a genetics clinic by virtue of having an unusual number of cases and are then perhaps further extended in a fashion that may or may not respect the rules of sequential sampling of pedigrees (15). Even if such further extension has been done systematically, the initial cluster may be difficult to define retrospectively as is necessary for ascertainment correction. Obviously, estimates of penetrance and allele frequency from such families will tend to be seriously overestimated unless the ascertainment and extension processes are taken into account, and, in most such instances, it is virtually impossible to characterize these processes in terms of any statistical sampling design. Nevertheless, it is theoretically possible to analyze the data in terms of the “retrospective” likelihood of the genotypes given the observed phenotypes [(13,18–19); Kraft P, Thomas DC: submitted for publication] to produce consistent estimates of population parameters; for genes with high penetration, such estimates based on linked markers (the “MOD score” approach) are generally thought to be fairly efficient but not for genes with low penetrance. A more fundamental problem, however, is that this approach implicitly assumes that the penetrance (or genetic relative risk) is homogeneous across families. If it is not, say, because of genetic heterogeneity due to other genes or environmental factors, then the average penetrance estimated from high-risk families will tend to be different from that estimated from a population-based series of families [(17,9); Kraft P, Thomas DC: submitted for publication]. For example, high-risk families may be segregating other (unknown) genes, some of whose effects could be incorrectly ascribed to the gene under study if this possibility is not explicitly addressed in the model. Further research is needed to study the potential biases and relative efficiency of designs based on high-risk families with the use of the retrospective likelihood approach.

**Case-Only Designs for Gene–Environment Interactions**

All of the above designs can be used for testing gene–environment interactions, as well as the main effect of a gene, as discussed by Goldstein and Andrieu (20). However, one additional design is available for this purpose that is not amenable to studying the main effects of genetic or environmental factors—the so-called “case-only” or “case–case” design (21–23). Under an assumption that genotype and an environment factor are independently distributed in the source population, one can interpret an association between genotype and exposure among cases as evidence of a gene–environment interaction. This design thus entails ascertainment of cases only and comparing the distribution of environmental factors between carriers and noncarriers.

**Statistical Analysis Issues**

Each of the above designs raises a host of statistical analysis issues that are generally beyond the scope of this paper. A few
key observations are worth pointing out, however. First, the analysis must respect the statistical sampling design, e.g., a matched case–control design would normally require a matched form of analysis (e.g., conditional logistic regression); a multi-stage sampling design must allow for the sampling fractions at each stage. Second, the form of analysis depends on the parameters of interest: A standard analysis of case–control data (whether using population or family controls) allows a direct estimate of the genetic relative risk and its modifiers, but estimation of absolute penetrances would require either external information (such as population rates) or some form of cohort design. Third, the bias and statistical efficiency of a design can only be defined in relation to a particular form of analysis, since some designs may allow more than one form of valid analysis. Finally, maximum likelihood analysis of a model that is correctly specified (both in terms of the true state of nature and the study design) will generally be asymptotically unbiased and fully efficient but may be more sensitive than other kinds of analysis to model misspecification. For example, it is likely that an analysis that ignores an important interaction or another important gene will generally lead to biased results whatever the design. Conversely, certain types of misspecification, such as residual familial aggregation, might be overcome by using robust “estimating equations” methods (24) or regressive models (25).

DISCUSSION

Our aim in the subsequent papers in this panel (4–7,13,20) is to provide some guidance about the best choice of design for addressing the various aims in the various contexts described at the outset. In the following section, we set up a general framework for addressing this problem and attempt to summarize what we believe is already known. To orient the reader to this discussion, we now lay out what we see as the “Big Questions.”

1) How can one determine whether a particular design will lead to unbiased estimates of a particular population parameter? For example, under what conditions might a case–control design with unrelated population controls lead to bias? In what ways might family-based, case–control or familial cohort studies be biased?

2) Are there important differences in power or statistical efficiency between designs? Is it possible to optimize the sampling plan for a particular objective?

3) Given a selection of a basic design, how should one decide whom to genotype?

4) What practical or other nonstatistical issues need to be taken into account?

5) What use, if any, can be made of families not selected in a population-based manner (e.g., heavily loaded pedigrees selected for linkage analysis) for estimating population parameters such as penetrance?

6) Once a gene has been identified, what are the best approaches to characterize other genes or environmental factors, or both?

7) Do the types of study depend on the types of gene involved, e.g., high versus low penetrance, common versus rare alleles?

8) Is it feasible to develop a population-based design that would be efficient both for characterization and for gene mapping studies?

Bias

Perhaps the question an investigator must consider first is whether to use unrelated or familial controls in any of the designs discussed above. To focus the discussion of this question, let us consider the case–control designs. Assuming the starting point is a population-based series of case patients, the primary concern about bias centers on the source of control subjects; both population control subjects and family control subjects are potentially subject to bias, but in quite different ways.

Population control subjects are particularly prone to a form of confounding bias known as “population stratification” (4), a situation that arises when both the gene frequency at the locus under study and the penetrances vary between subpopulations that are not accounted for by matching, stratification, or adjustment. Such confounding can arise through the effects of other genes, host factors, or environmental exposures. Suppose that the gene under study in fact has no causal effect but is associated in the population with some other gene that does have a causal effect; if the other gene is unknown or unmeasured, a spurious association with the gene under study will be seen. Such gene–gene associations can arise even in homogeneous populations through linkage disequilibrium with linked genes. Such associations are of considerable interest for the purpose of gene mapping but, from the perspective of gene characterization, are viewed as spurious. Gene–gene associations can also arise between unlinked genes in ethnically diverse populations simply as a result of a mixture of different allele frequencies at the two loci and are of no interest whatsoever. Likewise, it is possible for confounding by nongenetic risk factors to occur in a mixture of subpopulations with different prevalences of the two risk factors. As in any other form of epidemiologic study, confounding is controlled by matching or by stratified analysis, but these approaches require that the confounder be identifiable. Control of confounding can be very difficult for unknown genes that could have powerful risks and strong gradients between or even within ethnic groups, especially in freely mating multietnic populations. For example, it may be difficult to determine the relevant ethnic subgroup for a Caucasian subject (e.g., northern or southern European), and finding a suitable control subject for a case patient with grandparents from four different ethnic groups may be virtually impossible.

Several examples of spurious gene associations thought to be due to population stratification have been documented, such as the D2 allele of the dopamine receptor gene with alcoholism (26,27) and the Gm3,5,13,14 immunoglobulin haplotype with non-insulin-dependent diabetes among Pima-Papago Indians (28). Although these associations may be isolated examples, the magnitude of genetic risks and the frequent observation of large gradients between and within ethnic groups suggests that such confounding could be a much more serious problem than in other types of epidemiologic studies.

The problem of population stratification is completely eliminated by the use of siblings or pseudosibs as control subjects because they are descended from identical gene pools. Cousins do not provide such absolute control because of the possibility of intermarriage but, to the extent that people tend to marry within their ethnic groups, will generally avoid population stratification bias as well. Drawing cousin controls from both branches of mixed-race families can also help, but it can also introduce subtler biases (3).

Family control subjects are not immune to other forms of
bias, however. First, the pool of potential relative control subjects is limited and not every case may have a suitable control. If availability of a control is itself a risk factor (directly or indirectly, say by socioeconomic correlates of family size), then selection bias could result. Second, matching on age will generally be more difficult, again because of the limited number of relatives available. Control subjects should be required to have attained the case patient’s age at diagnosis and still be disease free, but this situation will generally lead to drawing control subjects from older relatives, leading to possible confounding by secular trends. Third, if multistage sampling is used to focus on multiple case families, then any restriction on case patients must apply equally to control subjects to avoid bias. This restriction is generally easy for sibling or pseudosib control subjects but not for cousins: If the case patient is required to have an affected mother or sibling, a sib control subject meets this requirement automatically, but a cousin would also have to have an affected mother or sibling with the same relationship to the case patient (i.e., if the case patient’s mother is the aunt of the case, the cousin’s mother would have to be the aunt of the case patient) (5).

Finally, we have alluded earlier to bias from model misspecification, in particular to analyses of familial cohort designs that ignore residual familial aggregation not accounted for by the measured genes or by other risk factors, or to phenotypic dependencies between family members. The standard conditional analysis of case–control studies avoids this difficulty by eliminating all stratum parameters (e.g., family- or ethnic-group specific baseline risks or allele frequency parameters), leading to unbiased estimates of the genetic relative risk. This option is not available for some types of familial cohort designs, and simply adding additional nuisance parameters to the model for each family is not a viable solution. A random-effects model may offer a valid analysis, but only if the form of that model is correctly specified (Kraft P, Thomas DC: submitted for publication).

Statistical Efficiency

By statistical efficiency, we refer to the precision of the parameter estimates from a particular study design and analysis combination. Generally, efficiency is described in terms of asymptotic relative efficiency (ARE) comparing the large sample variances of the estimated parameters from two alternative designs. The ARE can be interpreted as the ratio of sample sizes required to attain the same degree of statistical precision, and it is also closely related to the power of the test to reject a null hypothesis against some alternative. The large sample variances are computed from the inverse of the expected Fisher information (the negative of the inverse matrix of second derivatives of the log likelihood), averaged over all possible study outcomes (diseases and genotypes in a family) conditional on the study’s sampling plan and a particular set of model parameters. One must appreciate, however, that the ARE captures only the statistical sampling variability, not any of the other qualitative uncertainties or the relative costs of design alternatives.

As an example of relative efficiency calculations, consider the choice between use of population, sibling, pseudosib, or cousin control subjects in a case–control design for estimating relative risks. Without population heterogeneity (or matching on ethnicity), genotypes of unrelated controls would be independently distributed from those of the cases, whereas familial controls would tend to be associated with them, more closely for siblings than for cousins. This association would suggest that sibling control subjects would tend to be somewhat “overmatched” compared with population control subjects, leading to a less efficient design, whereas cousins would have intermediate efficiency. This intuition is correct, but the magnitude of the AREs depends on the true genetic model: For a multiplicative or a dominant gene, the sib control design has roughly 50% efficiency and the cousin control design 88% efficiency, relative to population control subjects; for a recessive gene with relative risk (RR) = 20 and allele frequency of 14%, these figures rise to 69% and 97%, respectively. The pseudosib design, however, is always at least as efficient as population controls (101%–106% for dominant and multiplicative models) and much more efficient (ARE = 231%) for major recessive genes (3,5).

Siblings (and to a lesser extent cousins) would also tend to have similar environmental exposures. Thus, one might expect that they would tend to be even less efficient for testing G × E interactions, but this intuition is not always correct (3). To test a twofold interaction RR, sib control subjects can be more efficient than any of the other four designs (ARE = 214%, 136%, and 275% for major recessive, multiplicative, and dominant genes, respectively), depending on such additional factors as the exposure prevalence (here 50%), the exposure concordance within families (odds ratio = 2), and the main effect of the environmental factor (RR = 2). The reason for this greater efficiency of sibling controls is that the most informative pairs are those that are genotype concordant and exposure discordant, and the proportion of such pairs is higher for sibling controls by virtue of their genetic relationship. Genotype-discordant, exposure-concordant, and jointly discordant pairs are also informative but are less common (3,5).

Such calculations are easily performed for a variety of design and analysis combinations and form the basis for power calculations for designing studies. Examples are discussed elsewhere in this volume for family-based, case–control (5), kin-cohort (7), and multistage (13) designs. To date, no general purpose software for design calculations presently exists, but this is an aim of the Informatics Center for the Cooperative Family Registries for Breast and Colorectal Cancer Research (Dr. Hoda Anton-Culver, principal investigator).

Practical Issues

The choice of a preferred design must, of course, turn on more than purely statistical considerations. We have already alluded to some of the potential difficulties in obtaining appropriate controls for a case–control study with either population or family controls—the difficulty of ethnic matching for the former, problems of availability and age-matching for familial controls, and family-history matching for multistage designs. Cooperation is another dimension: It has often been the experience of epidemiologists that family members are more highly motivated to participate than are randomly selected members of the population, but more difficult ethical issues arise when studying families. Families are frequently geographically dispersed, posing substantial logistical difficulties and additional costs. Comparisons of cost-efficiency must take account of the different types of costs—per family, per subject, per genotype, etc. The choice of design will also be influenced by such factors as survivorship: The pseudosib design is unlikely to be viable for such cancers as prostate because most parents will already be
dead, and the familial cohort designs with genotyping of family members may run into difficulties if most affected relatives are dead. The latter design also requires a carefully thought-out protocol for the selection of family members to genotype to avoid bias because of nonrandom selection and to maximize the information yield.

How Serious Is the Population Stratification Concern?

The potential bias because of confounding by ethnicity was discussed above and in greater detail by Caporaso et al. (4). As noted there, several textbook examples of spurious associations that appear to be because of population stratification have been documented, but the extent of the problem remains unknown. One other example of a different nature is worth mentioning. Insulin-dependent diabetes mellitus has consistently been found to be associated with a polymorphism (5'FP) near the 5' end of the insulin gene (29), but affected sib pair studies have found no evidence of linkage (30). However, the transmission-disequilibrium test (TDT), a variant of the pseudotest, was able to reject the joint null hypothesis of no association or no linkage (31). This situation appears to be an instance of a true population association that might have been incorrectly attributed to population stratification and could only be established by an appropriate family-control design. Here, it appears that the pure linkage tests failed because of their low power for detecting genes with only moderate effects.

How can we move beyond such anecdotal examples to a more systematic investigation of the population stratification problem? The difficulty is that truly causal genes for most diseases are probably still unknown, so we cannot directly observe the gene–gene associations that would resolve the issue. However, we do have ethnicity as a surrogate. Many diseases are known to have strong gradients (sometimes more than 100-fold) in incidence between ethnic groups, and many genes likewise show strong gradients in allele frequencies across ethnic groups. What is not known is how much of the gradient in incidence of a particular disease is directly attributable to gradients in frequency of already known genes and how much may be because of other factors (genetic or environmental) with which they may be associated. “Eccologic” studies with careful characterization of ethnic origin (as finely as possible, at least back to grandparents) might shed useful light on this question. Parallel family-based and population-based, case–control studies might also be useful, but they would require careful ethnic characterization to interpret any differences in conclusions. Studies might also be designed to take advantage of knowledge of allele frequencies in ancestral populations or to use a panel of polymorphic markers to infer ethnic origin (32,33).

What Are the Merits of Integrated Designs?

Zhao et al. (1,2) have suggested that it may be feasible to marry the objectives of gene discovery and gene characterization by a global research program firmly anchored in the principles of population-based, multistage sampling design for family studies. We have seen in this panel how these principles can indeed be put to good use to design unbiased and highly efficient studies for characterizing already identified genes. It is also evident that multistage sampling will allow heavily loaded pedigrees to be identified and extended in a manner that still allows estimation of population parameters. There are two outstanding questions, however. First, is this an efficient way to identify families that are highly informative for linkage analysis, compared with more anecdotal sources, such as high-risk cancer family clinics, and can any use be put to high-risk families ascertained in unsystematic ways for the purpose of estimating population parameters? Second, is it possible to develop optimal study designs for resources that are aimed at both purposes. For example, Whittemore and Halpern (16) note that the optimal designs for parameter estimation require all strata (defined by case–control status of the proband and family history) to be sampled with nonzero probabilities, whereas the optimal designs for linkage analysis (based on the yield of carrier families) entails enrolling all families from the highest prevalence strata until the desired sample size has been attained. These two objectives are thus fundamentally incompatible, and no design would be optimal for both purposes. The choice of a reasonable compromise would thus turn on the subjective importance of the two aims to the investigator. Further research along these lines would be helpful. Whittemore (34), Zhao et al. (2), and Haile et al. (14) in the Integration Panel of this volume address these questions further.

How Do We Decide Which Design to Use?

Several papers have addressed various aspects of these design trade-offs. In the author’s view, there are compelling advantages to the use of some form of family design for characterizing measured genes over case–control or cohort studies of independent individuals, most importantly because of their avoidance of the potential bias from population stratification. However, we recognize several difficulties with this approach, both statistical and practical. Within the class of family studies, there are a large number of choices, both in the selection of an overall study design as well as in the selection of specific family members to genotype. Although much is already known about the optimal choice of a sampling plan within a limited range of designs (e.g., choice of controls in case–control studies; choice of sampling fractions in multistage family cohort studies), there has been no systematic comparison of the full range of designs across the full spectrum of aims and contexts described at the outset. If we are ultimately to be able to address the question of what design is the best for a particular study, a much more systematic investigation of these design issues will be needed.

References

(9) Gail MH, Pee D, Benichou J, Carroll R. Designing studies to estimate the
penetrance of an identified autosomal dominant mutation: cohort, case–
(10) Whittemore AS. Logistic regression for family data. Biometrika 1995;82:
57–67.
et al. The risk of cancer associated with specific mutations of BRCA1 and
(12) Whittemore AS, Gong G, Intyre J. Prevalence and contribution of BRA1
mutations in breast cancer and ovarian cancer. Results from three U.S.
population-based case–control studies of ovarian cancer. Am J Hum Genet
1997;60:496–504.
(13) Siegmund KD, Whittemore AS, Thomas DC. Multistage sampling for dis-
(14) Haile RW, Siegmund KD, Gauderman WJ, Thomas DC. Study-design
issues in the development of the University of Southern California Con-
sortium’s Colorectal Cancer Family Registry. Monogr Natl Cancer Inst
(15) Cunnings C, Thompson EA. Ascertainment in the sequential sampling of
(17) Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort stud-
(18) Risch N. Segregation analysis incorporating linkage markers. I. Single-
locus models with an application to type I diabetes. Am J Hum Genet
1984;34:363–86.
(19) Hodge S, Elston R. Lods, wrods, and mods: the interpretation of lod scores
(20) Goldstein AM, Andrieu N. Detection of interaction involving identified
(21) Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models
and case-only designs for assessing susceptibility in population-based
(22) Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the
analysis of gene–environment interaction: case–control studies with no
(23) Umbach DM, Weinberg CR. Designing and analyzing case–control studies
to exploit independence of genotype and exposure. Stat Med 1997;16:
1731–43.
(24) Zhao LP, Hsu L, Holte S, Chen Y, Quiaoit F, Prentice RL. Combined
association and aggregation analysis of data from case–control family stud-
(25) Bonney G. Regressive models for familial and other binary traits. Biomet-
(26) Gelernter J, Goldman D, Risch N. The A1 allele at the D2 dopamine
P, et al. Allelic association of human dopamine D2 receptor gene in alco-
(28) Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm5,6,13,14 and type
2 diabetes mellitus: an association in American Indians with genetic ad-
(29) Bell GI, Horita S, Karam JH. A polymorphic locus near the human insulin
gene is associated with insulin-dependent diabetes mellitus. Diabetes 1984;
33:176–83.
(30) Spielman RS, Baur MP, Clerget-Darpoux F. Genetic analysis of IDDM: summ-
(31) Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage
disequilibrium: the insulin gene region and insulin-dependent diabetes mel-
Ethnic-affiliation estimation by use of population-specific DNA. Am J
Hum Genet 1997;60:957–64.
(33) Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect
population stratification in association studies. Am J Hum Genet 1999;65:
220–8.
(34) Whittemore AS, Nelson LM. Study design in genetic epidemiology: theo-
retical and practical considerations. Monogr Natl Cancer Inst 1999;26:
61–9.

**Note**

Supported in part by Public Health Service grants R01CA52862 (National Cancer Institute) and P30ES07048 (National Institute of Environmental Health Sciences), National Institutes of Health, Department of Health and Human Ser-

---

**Citation:** Journal of the National Cancer Institute Monographs No. 26, 1999 23